PROGNOSTIC VALUE OF HISTOLOGY AND LYMPH NODE STATUS IN BILHARZIASIS-BLADDER CANCER: OUTCOME PREDICTION USING NEURAL NETWORKS

W. Ji[1], R.N.G. Naguib[1], D. Petrovic[1], E. Gaura[1], and M.A. Ghoneim[2]
[1]BIOCORE, School of Mathematical and Information Sciences, Coventry University, Coventry, UK
[2]Urology and Nephrology Center, Mansoura University, Egypt

Abstract - **In this paper, the evaluation of two features in predicting the outcomes of patients with bilharziasis bladder cancer has been investigated using an RBF neural network. Prior to prediction, the feature subsets were extracted from the whole set of features for the purpose of providing a high performance of the network. Throughout the analysis of the prognostic feature combinations, two features, histological type and lymph node status, have been identified as the important indicators for outcome prediction of this type of cancer. The highest predictive accuracy reached 85.0% in this study.**

Keywords – **Schistosomiasis, feature extraction, classification, survival analysis, epidemiology**

I. INTRODUCTION

In the clinical issue of bladder cancer, histology type and lymph node status are two pathological markers that are always recorded. Based on histopathology, types of bladder cancer can be divided into two classes in general: transitional cell carcinoma (TCC) and non transitional cell carcinoma (non-TCC). The latter includes further four sub-classes: adenocarcinoma, squamous cell carcinoma, undifferentiated carcinoma, and mixed carcinoma. The TCC is present in 90% of bladder cancer cases, and has been investigated in many studies (e.g. [1]). The other non-TCC cases, although accounting for a small amount of the total cases (10%), express some specificity in risk and diagnosis factors. For instance, infections with members of the genus schistosoma are responsible for a high incidence of bladder cancer, 75% of which are squamous cell carcinomas [2]. The data set used in this study is collected from patients the majority of whom having bilharziasis infection history (accounts for 83.5% of the total cases). The investigation will focus on the sub-types of bladder cancer, namely TCC, squamous carcinoma, and adenocarcinoma.

Lymph Node status is used for staging bladder cancer within a TNM system (which considers tumour size, lymph node involvement, and distant metastasis). Some studies indicate that survival of cystectomy candidates with node positive bladder cancer is favourable when the primary tumour is confined to the bladder wall and lymph nodes involvement is minimal [3]. Therefore the relationship between lymph node status and disease progression needs to be addressed.

Identification of prognostic markers in relation to disease progression in cancer, generally, and bladder cancer, specifically, has long been a key issue of investigation. Recently reported studies aimed at identifying cancer prognostic markers proved that neural networks are capable of yielding a relatively high predictive accuracy with respect to an individual patient's course of disease progression. Amongst such neural techniques, the radial basis function algorithm (RBF), showed a high classification [4]. In the study by Qureshi et al, the RBF network was used to predict bladder cancer recurrence, stage progression and cancer-specific survival and was proved efficient [5]. Bilharziasis state was also identified as a significant prognostic marker in predicting the outcomes of patients with bilharziasis-associated bladder cancer [6].

In this study, we aim to determine the best feature subsets to improve predictive accuracy by means of the RBF network, and to assess the predictive ability of histology type and lymph node status by using the stratified feature subsets of the designed experimental data set.

II. METHODOLOGY

A. Data Set

The bladder cancer data set is described by eight clinical and pathological features which are shown in Table I. The total number of patients in this study is 321 with mean follow up of 5.0 years

TABLE I
FEATURES RECORDED IN THE BLADDER CANCER DATA SET

Features	Abbreviations
Histology	h
Tumour Grade	g
Lymph Nodes	l
Bilharziasis History	b
Stage	s
DNA Ploidy	d
Gender	e
Age Interval	a

The outcomes of the patients with bladder cancer to be considered for predictive analysis belong to two classes: alive free of disease and dead within five years of diagnosis. The distribution of these two classes in the data set is 178 for the former class and 143 for the latter.

B. Partition of the Data Set

The procedure developed in this study to assess the features contains four steps: (1) to reorganize the data set to allow characteristics of the features under consideration to be evaluated directly; (2) to extract useful features from the total set of features in order to improve predictive accuracy; (3) to use neural network results to identify significant markers; (4)

Report Documentation Page

Report Date	Report Type	Dates Covered (from... to)
25 Oct 2001	N/A	-

Title and Subtitle Prognostic Value of Histology and Lymph Node Status in Bilharziasis-Bladder Cancer: Outcome Prediction Using Neural Networks	**Contract Number**
	Grant Number
	Program Element Number
Author(s)	**Project Number**
	Task Number
	Work Unit Number
Performing Organization Name(s) and Address(es) BIOCORE School of Mathematical and Information Sciences Coventry University Coventry, UK	**Performing Organization Report Number**
Sponsoring/Monitoring Agency Name(s) and Address(es) US Army Research, Development & Standardization Group (UK) PSC 802 Box 15 FPO AE 09499-1500	**Sponsor/Monitor's Acronym(s)**
	Sponsor/Monitor's Report Number(s)

Distribution/Availability Statement
Approved for public release, distribution unlimited

Supplementary Notes
Papers from 23rd Annual International Conference of the IEEE Engineering in Medicine and Biology Society, October 25-28, 2001, held in Istanbul, Turkey. See also ADM001351 for entire conference on cd-rom.

Abstract

Subject Terms

Report Classification unclassified	**Classification of this page** unclassified
Classification of Abstract unclassified	**Limitation of Abstract** UU

Number of Pages
4

to analyse the marker combinations in order to find the prognostic value of the assessed features. The first step of this procedure is described as follows.

The original data set (ODS) includes all the patients used as a control experimental data set. When performing ODS prediction, the assessed features will be involved in, or omitted from, the feature subset to examine their predictive accuracy, in order to determine their prognostic ability. Meanwhile, another six sub-data sets are created from ODS in order to further determine the prognostic ability of the different features. The four data subsets for histology type are the TCC Data Set (TCCDS), the Squamous Data Set (SDS), the Adenocarcinoma Data Set (ADS), as well as others not specified in this study due to their corresponding small number of patients. The two data subsets for lymph node are the Lymph nodes Positive Data Set (LPDS) and the Lymph nodes Negative Data Set (LNDS). Table II shows the structure of these data subsets.

TABLE II
EXPERIMENTAL DATA SET STRUCTURE

Patients	ODS	TCCDS	SDS	ADS	others	LPDS	LNDS
Total No.	321	50	198	39	34	49	272
Survival	178	29	103	24	22	27	151
Mortality	143	21	95	15	12	22	121

C. The Radial Basis Function Network

The Radial Basis Function (RBF) network is a one hidden layer feedforward network in which the activation of the hidden layer is specified by a radial function. The output layer implements a linear combination of hidden neuron outputs as given by (1). The radial function used in this study is the Gaussian function $h_j(X)$ given by (2).

$$f(X) = \sum_{j=1}^{m} w_j h_j(X) \tag{1}$$

$$h_j(X) = \exp(-\frac{(X - c_j)^2}{r_j^2}) \tag{2}$$

where X is the input vector, w_j is the output layer weight vector, c_j and r_j are the center and radius of Gaussian function, and m is the number of hidden neurons.

When supervised learning is applied, the least squares algorithm is used to obtain the output layer weights. This is referred to as the normal equation:

$$W = (H^T H + \Lambda)^{-1} H^T Y \tag{3}$$

where H is the design matrix composed by $h_j(X)$, Y is the vector of training set outputs, and Λ is the regularisation parameter vector (here set at 0).

The key issue in an RBF network design is to select the number of hidden neurons and their parameters. In this study, the K-means cluster technique [7] was adopted to produce the initial size of the hidden layer and the parameters, then a pruning technique is applied according to the RBF validation procedure.

D. Predictive Analysis and Feature Assessment

An efficient way to improve the performance of a neural network is to provide it with representative information. This therefore implies undertaking feature extraction. Appropriate feature selection helps to facilitate classification by eliminating noisy or non-representative features that can impede the prediction. In this study, and following feature selection, two issues are being addressed: (1) whether the assessed feature involved in the best predictive feature subsets, and (2) whether the feature subsets are the same for the different experimental data sets (i.e. ODS, TCCDS, SDS, ADS for histology and LPDS, LNDS for lymph nodes). The procedure employed in this study consists of two steps: (a) to present only one feature each time to the RBF network, (b) to present all features, except one, each time.

The overall process including the splitting of the experimental data subset, the feature selection scheme and the RBF algorithm have been implemented using the Clementine data mining environment [8]. The stream conducted by visual programming provides the comparative evaluation of each feature subset on prediction. During the training period, the alternation of the cluster number which determines the radial function center and the size of the hidden layer has been shown to have an apparent impact on the overall predictive accuracy.

III. RESULTS

The predictive results are focused on two aspects: (1) predictive accuracy for the total set of patients of every experimental data set (TEPA) and for the two classes of outcomes (PAS represents the predictive accuracy on survival patients and PAD on those who died), and (2) the sensitivity and specificity (represented by Sen. and Spec.) of the outcomes. The analysis is based on the feature combinations stratified from the entire set of features using the RBF network. This is shown in the first column of the prediction tables (Tables III to IX, the subsets shown in the tables are part of the predictive results which have the relatively high predictive accuracy).

Table III indicates that the predictive accuracy is 62.7% when all features are involved, while the other four parameters are 67.1%, 57.9%, 64.0%, and 61.1%. The best result with the comprehensive consideration for each evaluation parameter is observed to be 66.5%, 69.7%, 62.5%, 69.66%, and 62.5% in the set of {h, g, l, b, d}.

Before assessing prediction on the TCCDS, SDS, and ADS separately, the predictive results were found by putting all three data sets together (i.e. Involving Three Data Sets, ITDS), see table IV.

Using the results in Tables IV to VII, in considering the disease histology type, the best TEPA is found to be 70.1% for the feature subset of {g,l,b,s,d,a}for prediction of the three types of bladder cancer (ITDS, Table IV). Also, the PAS reaches a higher degree of accuracy (82.1%) than the prediction without feature selection (first row in table IV).

When the data sets are split (Tables V, VI, and VII), the predictive accuracy is always higher than in the case when they are used together (ITDS, Table IV). Amongst those, the prediction of ADS shows the highest degree of accuracy compared with the other two types of cancers (TCCDS and SDS). The TEPA for these three data sets are 85.0% for the feature subset {g,s,d,l} for ADS, 76.0% for the subset {g,s,b,a} for TCCDS, and 71.7% for the subset {g,l,b,s,d} for SDS. The highest PAS even reaches 100% for the ADS prediction (Table VII).

TABLE III
RBF PREDICTION ON ODS

Subsets	TEPA(%)	Sen.(%)	Spec.(%)	PAS(%)	PAD(%)
{h,g,l,b,s,d,e,a}	62.7	67.1	57.9	64.0	61.1
{h,g,l,b,s,d}	65.2	68.1	61.4	69.7	59.7
{h,g,l,b,d}	66.5	69.7	62.5	69.7	62.5
{g,b,s,d}	65.8	68.5	62.3	70.8	59.7

TABLE IV
RBF PREDICTION ON ITDS

Subsets	TEPA(%)	Sen.(%)	Spec.(%)	PAS(%)	PAD(%)
{g,l,b,s,d,e,a}	63.9	65.1	62.1	71.8	54.6
{g,l,b,s,d,a}	70.1	68.8	72.6	82.1	56.1
{g,l,b,s,d,e}	66.7	66.0	68.0	79.5	51.5
{g,s,d,e,a}	68.1	67.4	69.2	79.5	54.6

TABLE V
RBF PREDICTION ON TCCDS

Subsets	TEPA(%)	Sen.(%)	Spec.(%)	PAS(%)	PAD(%)
{g,l,b,s,d,e,a}	72.2	73.3	70.0	78.6	63.6
{g,l,b,s,e,a}	76.0	78.6	72.7	78.6	72.7
{g,b,d,e,a}	76.0	78.6	72.7	78.6	72.7
{g,s,b,a}	76.0	78.6	72.7	78.6	72.7

TABLE VI
RBF PREDICTION ON SDS

Subsets	TEPA(%)	Sen.(%)	Spec.(%)	PAS(%)	PAD(%)
{g,l,b,s,d,e,a}	65.7	66.0	65.2	68.6	62.5
{g,l,s,d,e,a)	68.7	67.9	69.8	74.5	62.5
{g,l,b,s,d}	71.1	72.6	70.8	72.6	70.8
{g,l,s,d,a}	68.7	69.2	68.1	70.6	66.7

TABLE VII
RBF PREDICTION ON ADS

Subsets	TEPA(%)	Sen.(%)	Spec.(%)	PAS(%)	PAD(%)
{g,l,b,s,d,e,a}	80.0	78.6	83.3	91.7	62.5
{g,b,s,d,e,a}	80.0	78.6	83.3	91.7	62.5
{g,l,b,s,d}	85.0	80.0	100.0	100.0	62.5
{g,l,s,d}	85.0	80.0	100.0	100.0	62.5

TABLE VIII
RBF PREDICTION ON LPDS

Subsets	TEPA(%)	Sen.(%)	Spec.(%)	PAS(%)	PAD(%)
{h,g,b,s,d,e,a}	72.0	73.3	70.0	78.6	63.6
{h,g,s,d,e,a}	68.0	71.4	63.6	71.4	63.6
{h,g,b,s,d,e}	72.0	76.9	66.8	71.4	72.7
{h,g,s,d,e}	68.0	75.0	61.5	64.3	72.7

In the evaluation of lymph node status, some subsets of the features also express a distinctly higher performance than when all features are involved (Tables VIII and IX). In the LPDS table (Table VIII), the highest predictive accuracy reaches 72.0% and the PAD attains 72.7% for the feature subset {h,g,b,s,d,e}. In the LNDS table (Table IX), the feature subset {h,b,s,d,e} provides the best performance characteristics with a predictive accuracy of 70.6%, sensitivity 77.8%, and specificity 64.4%.

The prognostic marker combinations stratified from the prediction tables for the experimental data sets are summarised in Tables X and XI, showing clearly the relations within the markers, which will be discussed in more detail in the next section.

TABLE IX
RBF PREDICTION ON LNDS

Subsets	TEPA(%)	Sen.(%)	Spec.(%)	PAS(%)	PAD(%)
{h,g,b,s,d,e,a}	66.9	70.8	62.5	68.0	65.6
{h,b,s,d,e}	70.6	77.8	64.4	65.3	77.1
{h,g,d,e,a}	69.1	69.4	68.6	78.7	57.4
{h,b,s,d}	69.1	80.0	61.7	58.7	82.0

TABLE X
MARKER COMBINATIONS FOR DIFFERENT TYPES OF TUMOURS

Feature	g	l	b	d	s	e	a
TCC	*		*		*		*
Squa.	*	*	*	*	*		
Aden.	*	*		*	*		
IT[a]	*	*	*	*	*		*

[a]IT: Combining the three types together

TABLE XI
MARKER COMBINATIONS FOR DIFFERENT POLARITY OF LYMPH NODES

Feature	h	g	b	d	s	e	a
LPDS	*	*	*	*	*	*	
LNDS	*		*	*	*	*	

IV. DISCUSSION

All of the predictions, either on the aggregate or separate data, have been conducted with both the whole feature sets and feature subsets. Therefore, it is necessary to separately discuss the following: the overall predictive analysis which focuses on the performance of TEPA, the assessment of lymph nodes, and the assessment of histology type.

A. Overall Predictive Accuracy

Results obtained demonstrate that the feature subset prediction can improve the overall accuracy compared with the presentation of the entire features using the RBF network. From Tables III to IX, the following observations can be made:

1) For ODS analysis, omission of any one feature leads to a decrease in accuracy (although not presented here). However, there is noticeable interaction amongst the features, omission of two or three features may lead to an increase in predictive accuracy (Table III). This demonstrates that the interaction which exists within features greatly influences prediction.
2) For the LPDS and LNDS cases, the overall predictive accuracy increased more significantly than that of ODS with the feature subsets prediction. This implies that the existence, or not, of lymph node status information has a definite impact on either one of the parameters

(sensitivity or specificity) despite the possibly high predictive accuracy achieved.

3) The subset {g,b,s} has a significant impact on prediction of the patients who are alive since the PAS reaches 91.1% in the data set of ODS.
4) By analysing the results from the four data sets for histology type, it is observed that prediction within a single type of tumour can attain a higher accuracy than the prediction involving all three types.
5) By comparing the results from the two feature data sets (Tables IV, V, VI, and VII with VIII, IX, and X), although feature subsets prediction within the lymph nodes status data sets yields a higher predictive accuracy than in the case of ODS, the histology data sets can all produce a better prediction. This fact demonstrates that histology itself is an active indicator of patient survival.
6) Referring to ODS again, omission of histology yielded a lower predictive accuracy than most of the other single feature omission cases and the best feature subset prediction also involved this marker. This result supports the conclusion that histology is one of the most important indicators of bladder cancer outcome prediction.

B. Assessment of Histology

From Tables IV to VII, the following observations could be made:

1) Almost all feature combinations yielding high predictive accuracy listed in Tables V, VI, and VII discarded the same feature of gender. This demonstrates that gender is not a prognostic marker for any of the three single types of the bladder cancer. This conclusion has been also supported by the fact that the best prognostic feature combination using the three types of data sets (ITDS, table IV) is the one excluding the feature of gender.
2) Although for TCC prediction, the prognostic combination involving age yielded the highest degree of accuracy, it is not involved in the prediction combination for the other two types of bladder cancer. Therefore, the factor of age was not used as a prognostic marker for the total set of patients (ITDS, Table IV). This indicates that the age of the patients does not influence the outcomes of one's disease in the case of non-TCC cancers, but will have a significant impact on the outcome prediction of the patients with TCC.
3) The three types of histology experimental sets show a distinct level of prediction. To predict outcome for adenocarcinoma patients is the easiest, since most of the TEPA parameters are $\geq$ 80% whilst the highest one reaches 85%. But in predicting squamous carcinoma, the TEPA can only reach 71.7%. This implies that division into separate types of tumours can significantly improve the predictive performance.
4) Considering the two outcomes (survival and death) in Tables V, VI, and VII, it is observed that almost all prediction with the proper feature combination gave a balanced value of sensitivity and specificity (data set TCCDS, SDS, and ADS), except for the case of adenocarcinoma. The prediction of ADS is more biased toward survival than death (specificity even reaches 100% using the feature subset {g,s,d,l}).

C. Assessment of Lymph Node Status

Lymph node status is considered to have a direct relation to the spread of the cancer [3]. In this study, the major observations can be made as below:

1) It is obvious that for both data sets, we can always find suitable feature subsets to obtain a high predictive accuracy. That implies that the feature subset prediction technique is also efficient for LPDS and LNDS.
2) The prediction accuracy within patients whose lymph nodes test positive is higher than for negative lymph node patients.
3) Although the prognostic combinations in both data sets are different, both of them omitted the feature of *Age*. This demonstrates that this feature is not a significant prognostic marker.
4) The identified prognostic markers for the two data sets (Table XI) are very similar, with LPDS prediction including only one more marker involved than LNDS. This marker (Tumour Grade) is shown to be significant in predicting survival and death for lymph node positive patients.

V. CONCLUSIONS

The proposed analyses have shown the ability to assess two markers, lymph node status and disease histology, based on the prediction of the RBF neural network. From this study, the important combination for ODS prediction is the subset {h,g,l,b,d}. Here, both of the assessed features, lymph node status and disease histology, have been evaluated to be important prognostic markers in the predictive analysis. For each experimental set, the prognostic markers are {h,g,b,s,d,e} for LPDS, {h,b,s,d,e} for LNDS, {g,l,b,s,d,a} for ITDS, {g,b,s,a} for TCCDS, {g,l,b,s,d} for SDS, and {g,l,s,d} for ADS.

REFERENCES

[1] M.P. Raitanen, and T.L.J. Tammela, "Relationship between blood-group and tumour grade, number, size, stage, recurrence and survival in patients with transitional-cell carcinoma of the bladder," *Scand J Urol Nephrol*, vol. 27, no. 3, pp. 343-347, 1993.

[2] S.L. Johansson, and S.M. Cohen, "Epidemiology and etiology of bladder cancer," *Semin Surg Oncol*, vol. 13, no. 5, pp. 291-298, 1997.

[3] J. Vieweg, J.E. Gschwend, H.W. Herr, and W.R. Fair, "The impact of primary stage on survival in patients with lymph node positive bladder cancer," *J Urol*, vol. 161, no. 1, pp. 72-76, 1999.

[4] Y.S. Hwang. and S.Y. Bang, "An efficient method to construct a radial basis function neural network classifier," *Neural Networks*, vol. 10, no. 8, pp. 1495-1503, 1997.

[5] K.N. Qureshi., R.N.G. Naguib, F.C. Hamdy, D.E. Neal, and J.K. Mellon, "Neural network analysis of clinicopathological and molecular markers in bladder cancer," *J Urol*, vol. 163, no. 2, pp. 630-633, 2000.

[6] W. Ji, S. Tahzib, R.N.G. Naguib, and M. Ghoneim, "Identification of significant prognostic markers for outcome prediction in bilharziasis-associated bladder cancer," *Proceedings of the World Congress on Medical Physic and Biomedical Engineering*, Chicago July, 2000.

[7] J.C. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms,* Plenum Press, New York, pp. 55-58, 1981.

[8] *Clementine User Guide Version 5*, Integral Solutions, 1998.